DOCUMENT RESUME

ED 036 521 | 24 | TE 001 727

AUTHOR	ALLEN, R. R.; AND OTHERS
TITLE	THE DEVELOPMENT OF THE WISCONSIN TESTS OF TESTIMONY AND REASONING ASSESSMENT (WISTTRA). REPORT FROM THE CONCEPTS IN VERBAL ARGUMENT PROJECT.
INSTITUTION	WISCONSIN UNIV., MADISON. RESEARCH AND DEVELOPMENT CENTER FOR COGNITIVE LEARNING.
SPONS AGENCY	OFFICE OF EDUCATION (DHEW), WASHINGTON, D.C. BUREAU OF RESEARCH.
REPORT NO	TR-80
BUREAU NO	BR-5-0216
PUB DATE	APR 69
CONTRACT	OEC-5-10-154
NOTE	44P.
EDRS PRICE	EDRS PRICE MF-$0.25 HC-$2.30
DESCRIPTORS	COGNITIVE MEASUREMENT, COGNITIVE TESTS, *CRITICAL THINKING, *EDUCATIONAL TESTING, EVALUATION CRITERIA, EVALUATION METHODS, EVALUATION TECHNIQUES, *EVALUATIVE THINKING, MENTAL TESTS, PRODUCTIVE THINKING, STUDENT TESTING, *TEST CONSTRUCTION, TEST INTERPRETATION, TEST RELIABILITY, TEST VALIDITY, *VERBAL ABILITY, VERBAL TESTS
IDENTIFIERS	*WISCONSIN TESTS TESTIMONY REASONING ASSESSMENT, WISTTRA

ABSTRACT

THIS REPORT PRESENTS AN OVERVIEW OF RESEARCH RELATED TO THE DEVELOPMENT OF THE "WISCONSIN TESTS OF TESTIMONY AND REASONING ASSESSMENT," A BATTERY OF SEVEN TESTS FOR ASSESSING STUDENT DEVELOPMENT IN THE MASTERY OF RELEVANT CONCEPTS AND SKILLS OF VERBAL ARGUMENT FOR GRADES 10-12. PROVIDED ARE (1) A DISCUSSION OF THE RATIONALE FOR AND PURPOSES OF THE TESTS, (2) A DISCUSSION OF THE DEVELOPMENT OF THE TESTS, AND (3) A DESCRIPTION OF EACH TEST, RELIABILITY AND ITEM DATA, AND A BRIEF DISCUSSION OF THE SIGNIFICANCE OF THE DATA GATHERED BY ADMINISTERING THE FIFTH EDITION OF THE TEST BATTERY TO OVER 3,000 JUNIOR/SENIOR HIGH STUDENTS IN FOUR WISCONSIN SCHOOL SYSTEMS. INCLUDED IS A DISCUSSION OF THE SPECIFIC OBJECTIVES OF THE TESTS: TO MEASURE THE STUDENT'S ABILITY TO EVALUATE TESTIMONY IN TERMS OF INTERNAL CRITERIA, CONSISTENCY WITH OTHER TESTIMONY, AND RECENCY AND PROXIMITY; TO RECOGNIZE AND SELECT WARRANTS IN ARGUMENTS; TO RECOGNIZE STATEMENTS WHICH ANSWER RESERVATIONS IN ARGUMENTS; TO SELECT RESERVATIONS IN ARGUMENTS; AND TO SELECT CLAIMS IN ARGUMENTS. THIRTY-ONE TABLES AND TWO FIGURES OFFER FURTHER INFORMATION. (AUTHOR/LH)

U.S. DEPARTMENT OF HEALTH, EDUCATION & WELFARE
OFFICE OF EDUCATION
THIS DOCUMENT HAS BEEN REPRODUCED EXACTLY AS RECEIVED FROM THE PERSON OR ORGANIZATION ORIGINATING IT. POINTS OF VIEW OR OPINIONS STATED DO NOT NECESSARILY REPRESENT OFFICIAL OFFICE OF EDUCATION POSITION OR POLICY.

BR-5-0216
TR 80
PA-24
OE BR

ED036521

Technical Report No. 80

THE DEVELOPMENT OF THE WISCONSIN TESTS OF TESTIMONY AND REASONING ASSESSMENT (WISTTRA)

By R. R. Allen, Jerry D. Feezel, Fred J. Kauffeld, and Margaret L. Harris

Report from the Concepts in Verbal Argument Project
R. R. Allen, Principal Investigator

Wisconsin Research and Development
Center for Cognitive Learning
The University of Wisconsin
Madison, Wisconsin

April 1969

TE001727

The research reported herein was performed pursuant to a contract with the United States Office of Education, Department of Health, Education, and Welfare, under the provisions of the Cooperative Research Program. The opinions expressed in this publication do not necessarily reflect the position or policy of the Office of Education and no official endorsement by the Office of Education should be inferred.

Center No. C-03 / Contract OE 5-10-154

NATIONAL EVALUATION COMMITTEE

Samuel Brownell
Professor of Urban Education
Graduate School
Yale University

Henry Chauncey
President
Educational Testing Service

Elizabeth Koontz
President
National Education Association

Patrick Suppes
Professor
Department of Mathematics
Stanford University

Launor F. Carter
Senior Vice President on Technology and Development
System Development Corporation

Martin Deutsch
Director, Institute for Developmental Studies
New York Medical College

Roderick McPhee
President
Punahou School, Honolulu

***Benton J. Underwood**
Professor
Department of Psychology
Northwestern University

Francis S. Chase
Professor
Department of Education
University of Chicago

Jack Edling
Director, Teaching Research Division
Oregon State System of Higher Education

G. Wesley Sowards
Director, Elementary Education
Florida State University

UNIVERSITY POLICY REVIEW BOARD

Leonard Berkowitz
Chairman
Department of Psychology

John Guy Fowlkes
Director
Wisconsin Improvement Program

Herbert J. Klausmeier
Director, R & D Center
Professor of Educational Psychology

M. Crawford Young
Associate Dean
The Graduate School

Archie A. Buchmiller
Deputy State Superintendent
Department of Public Instruction

Robert E. Grinder
Chairman
Department of Educational Psychology

Donald J. McCarty
Dean
School of Education

***James W. Cleary**
Vice Chancellor for Academic Affairs

H. Clifton Hutchins
Chairman
Department of Curriculum and Instruction

Ira Sharkansky
Associate Professor of Political Science

Leon D. Epstein
Dean
College of Letters and Science

Clauston Jenkins
Assistant Director
Coordinating Committee for Higher Education

Henry C. Weinlick
Executive Secretary
Wisconsin Education Association

EXECUTIVE COMMITTEE

Edgar F. Borgatta
Brittingham Professor of Sociology

Russell J. Hosler
Professor of Curriculum and Instruction and of Business

Wayne Otto
Professor of Curriculum and Instruction (Reading)

Richard L. Venezky
Assistant Professor of English and of Computer Sciences

Max R. Goodson
Professor of Educational Policy Studies

***Herbert J. Klausmeier**
Director, R & D Center
Professor of Educational Psychology

Robert G. Petzold
Associate Dean of the School of Education
Professor of Curriculum and Instruction and of Music

FACULTY OF PRINCIPAL INVESTIGATORS

Ronald R. Allen
Associate Professor of Speech and of Curriculum and Instruction

Gary A. Davis
Associate Professor of Educational Psychology

Max R. Goodson
Professor of Educational Policy Studies

Richard G. Morrow
Assistant Professor of Educational Administration

Vernon L. Allen
Associate Professor of Psychology
(On leave 1968-69)

M. Vere DeVault
Professor of Curriculum and Instruction (Mathematics)

Warren O. Hagstrom
Professor of Sociology

Wayne Otto
Professor of Curriculum and Instruction (Reading)

Nathan S. Blount
Associate Professor of English and of Curriculum and Instruction

Frank H. Farley
Assistant Professor of Educational Psychology

John G. Harvey
Associate Professor of Mathematics and Curriculum and Instruction

Milton O. Pella
Professor of Curriculum and Instruction (Science)

Robert C. Calfee
Associate Professor of Psychology

John Guy Fowlkes (Advisor)
Professor of Educational Administration
Director of the Wisconsin Improvement Program

Herbert J. Klausmeier
Director, R & D Center
Professor of Educational Psychology

Thomas A. Romberg
Assistant Professor of Mathematics and of Curriculum and Instruction

Robert E. Davidson
Assistant Professor of Educational Psychology

Lester S. Golub
Lecturer in Curriculum and Instruction and in English

Burton W. Kreitlow
Professor of Educational Policy Studies and of Agricultural and Extension Education

Richard L. Venezky
Assistant Professor of English and of Computer Sciences

MANAGEMENT COUNCIL

***Herbert J. Klausmeier**
Director, R & D Center
Acting Director, Program 1

Thomas A. Romberg
Director
Programs 2 and 3

James E. Walter
Director
Dissemination Section

Dan G. Woolpert
Director
Operations and Business

Mary R. Quilling
Director
Technical Section

* COMMITTEE CHAIRMAN

STATEMENT OF FOCUS

The Wisconsin Research and Development Center for Cognitive Learning focuses on contributing to a better understanding of cognitive learning by children and youth and to the improvement of related educational practices. The strategy for research and development is comprehensive. It includes basic research to generate new knowledge about the conditions and processes of learning and about the processes of instruction, and the subsequent development of research-based instructional materials, many of which are designed for use by teachers and others for use by students. These materials are tested and refined in school settings. Throughout these operations behavioral scientists, curriculum experts, academic scholars, and school people interact, insuring that the results of Center activities are based soundly on knowledge of subject matter and cognitive learning and that they are applied to the improvement of educational practice.

This Technical Report is from the Concepts in Verbal Argument Project in Program 2. General objectives of the Program are to establish rationale and strategy for developing instructional systems, to identify sequences of concepts and cognitive skills, to develop assessment procedures for those concepts and skills, to identify or develop instructional materials associated with the concepts and cognitive skills, and to generate new knowledge about instructional procedures. Contributing to these Program objectives, the staff of the project developed a semiprogramed course in verbal argument and related tests for use at the high school level. The project staff prepared the materials on the basis of an outline of concepts and critical skills developed from an evaluation of everyday discourse.

CONTENTS

		Page
	List of Tables and Figures	vii
	Abstract	ix
I	Introduction	1
II	Rationale and Purposes	2
III	Development of the Tests	4
	Editions	4
	Test Instructions	4
	Test Vocabulary	4
	Student Interest	4
	Subject Matter of Items	5
	Content Validity	6
	Internal Consistency Reliability	7
	Item Characteristics	7
	Summary	8
IV	Discussion of Specific Tests	10
	Testimony I: Appraising Testimony in Terms of Internal Criteria	10
	Testimony II: Appraising Testimony in Terms of External Criteria: Consistency with Other Testimony	10
	Testimony III: Appraising Testimony in Terms of External Criteria: Recency and Proximity	11
	Reasoning I: Recognizing and Selecting Warrants in Arguments	11
	Reasoning II: Recognizing Statements Which Answer Reservations in Arguments	12
	Reasoning III: Selecting Reservations in Arguments	12
	Reasoning IV: Selecting Claims in Arguments	13
V	Reliability Estimates and Item Statistics	14
	Sample	14
	Reliability Estimates	14
	Item Statistics	14
VI	Conclusions	36
	References	37

LIST OF TABLES AND FIGURES

Table		Page
1	Test Administrations in the Development of WISTTRA	5
2	Panel of Student Consultants	5
3	Preferable Difficulty Levels for WISTTRA	8
4	Testimony I: Reliability Estimates for the Total Test	15
5	Testimony I: Reliability Estimates for the Accept Subtest	15
6	Testimony I: Reliability Estimates for the Bias Subtest	16
7	Testimony I: Reliability Estimates for the Position Subtest	16
8	Testimony I: Reliability Estimates for the Competence Subtest	17
9	Testimony I: Reliability Estimates for the Qualification Subtest	17
10	Testimony II: Reliability Estimates	18
11	Testimony III: Reliability Estimates for the Total Test	18
12	Testimony III: Reliability Estimates for the Recency Subtest	19
13	Testimony III: Reliability Estimates for the Proximity Subtest	19
14	Reasoning I: Reliability Estimates	20
15	Reasoning II: Reliability Estimates	20
16	Reasoning III: Reliability Estimates	21
17	Reasoning IV: Reliability Estimates	21
18	Testimony I: Item Statistics for the Total Test	22
19	Testimony I: Item Statistics for the Accept Subtest	23
20	Testimony I: Item Statistics for the Bias Subtest	24
21	Testimony I: Item Statistics for the Position Subtest	25

Table

		Page
22	Testimony I: Item Statistics for the Competence Subtest	26
23	Testimony I: Item Statistics for the Qualification Subtest	27
24	Testimony II: Item Statistics	28
25	Testimony III: Item Statistics for the Total Test	29
26	Testimony III: Item Statistics for the Recency Subtest	30
27	Testimony III: Item Statistics for the Proximity Subtest	31
28	Reasoning I: Item Statistics	32
29	Reasoning II: Item Statistics	33
30	Reasoning III: Item Statistics	34
31	Reasoning IV: Item Statistics	35

Figure

1	Relationship of WISTTRA to Concepts Identified	6
2	A Typical Item Characteristic Curve	9

ABSTRACT

This paper reports the development of a test battery for measuring student mastery of certain verbal skills basic to critical thinking. The battery, collectively entitled The Wisconsin Tests of Testimony and Reasoning Assessment, consists of three tests related to testimony and four tests related to arguments developed through reasoning. The basic rationale for the tests and major considerations in test development are explained. Each test is described and appropriate reliability estimates and item statistics are presented. The particular data presented were gathered by an administration of the fifth edition of the test battery to over 3,000 junior/senior high students in four Wisconsin school systems.

I
INTRODUCTION

This report presents an overview of research related to the development of a test battery for measuring student mastery of certain verbal skills basic to critical thinking. The report provides (1) a discussion of the rationale for and purposes of the tests, (2) a discussion of the development of the tests, and (3) a discussion of each test including a description of the test, reliability and item data, and a brief discussion of that data. The following discussion, then, is intended strictly as a report of test development; it is not offered as a guide to the use or interpretation of the tests.

II
RATIONALE AND PURPOSES

In developing the Wisconsin Tests of Testimony and Reasoning Assessment (WISTTRA) the researchers sought (1) to create a valid and reliable testing instrument to be generally available for assessing student development in the mastery of relevant concepts and skills of verbal argument for Grades 10–12, and (2) to gather data from administrations of this battery useful in the development of related instructional materials. The tests have been published by and a sample copy is available from the Wisconsin Research and Development Center for Cognitive Learning (Allen, Feezel, & Kauffeld, 1968). For a more complete statement of the project's rationale the reader should consult A Taxonomy of Concepts and Critical Abilities Related to the Evaluation of Verbal Argument (Allen, Feezel, & Kauffeld, 1967).

WISTTRA is based on a view of verbal argument articulated by the English philosopher Stephen Toulmin (1958) and adapted to the field of ordinary argument by the investigators. His program, an off-shoot of Rylean language philosophy, is developed on two central points: (1) the habits of reasoning utilized in any field of inquiry involve rules for evaluating inferences much richer than the field-invariant schemes worked out by formal logicians, and (2) an adequate account of such rules can only be worked out by attending to the nature of particular fields of inquiry. The tests discussed here grew out of a definition of rules of inference fundamentally important to the field of ordinary, i.e., nontechnical, argument.

In order to characterize the rules of inference appropriate to ordinary arguments the researchers first isolated three major requirements imposed by the nature of plain discourse on ordinary reasoning: (1) ordinary arguments must be able to take the reports of other people (testimony) as an important source of primary data, (2) ordinary arguments must be able to provide reasons for a wide variety of claim types, and (3) ordinary arguments must be able to handle inferences which utilize categories built on multidimensional, loosely related configurations of criteria.

On this foundation two orders of concepts were distinguished: (1) those related to appraising the testimony of others and (2) those related to appraising the strength of reasons given for a claim. Concepts used in appraising testimony may be grouped into two clusters: (a) internal tests—position to observe, ability to observe, bias, and qualification for judging —and (b) external tests—primary as compared with secondary information, recent as compared with dated information, and consistent as compared with inconsistent information. Concepts used in assessing the strength of reasons may be grouped into two clusters: (a) those related to the structure of arguments—data, warrant, claim, and reservation—and (b) rules for assessing arguments developed through reasoning. The latter are sensitive to the type of argument being assessed and are represented in the test battery as they are used to assess sign, class, causal, alternative, parallel case, comparative, and warrant supportive arguments. Basic skills used in assessing arguments developed through reasoning include (1) the ability to detect missing parts of an argument, (2) the ability to discern the relevance of objections, and (3) the ability to recognize appropriate conclusions.

The researchers saw a compelling need for a test battery adapted to just this configuration of concepts and skills, because measuring instruments developed on the assumption that rules of inference are field-invariant do not assess the student's mastery of the skills and concepts appropriate to ordinary argument. Commonly, tests based on field-invariant logics simply measure the student's mastery of the rules of inference appropriate to some preferred field of specialized inquiry. Such tests are useful when information about the student's mastery of the reasoning habits of the preferred

field are of interest, but they may give a very distorted picture of a student's ability to handle everyday arguments.

In particular, tests of reasoning based on field-invariant logics usually neglect the concepts and skills related to assessing testimony and discerning the relevance of an objection. Tests based on the highly mechanical procedures for induction and deduction prescribed by type logics are particularly vulnerable to this criticism. Few ordinary arguments involve questions which can be resolved by direct observations of the participants, and still fewer involve questions which can be fully analyzed against the tidy categories such systems require. WISTTRA was developed to assess the student's ability to evaluate adequacy of testimony and to recognize the structure that is present in ordinary arguments and raise pertinent objections based on the rules of inference appropriate to that structure.

III
DEVELOPMENT OF THE TESTS

EDITIONS

The development of WISTTRA constitutes one phase of research related to the development of student abilities in the assessment of verbal arguments. From the project's inception the researchers recognized the need for an appropriate testing instrument. Work on the tests was begun in February of 1966 and continued through April of 1968. During that period the instrument went through four experimental editions in which its focus was narrowed from Grades 7–12 to Grades 10–12 and its items analyzed and revised for greater precision and reliability. During that period portions of the battery were pretested on four occasions and a normative study was conducted with a fifth edition of the battery (See Table 1.)

Development of the tests will be discussed in terms of the considerations and criteria which the investigators used in decisions related to test instructions, test vocabulary, student interest, subject matter of items, content validity, internal-consistency reliability, and item characteristics.

TEST INSTRUCTIONS

As is often the case, drawing up instructions for the various tests in the battery required balancing the need to provide sufficient information to complete the task against the demand that test instructions not teach the student skills the test seeks to measure. In order to minimize the confounding effects of test instructions, two forms of instructions were used in Pretest One. The two forms differed only in that Form A included an example of the response task while Form B did not. The two instructional forms did not yield significant differences in student responses, but from general indications and conversations with students the longer form was selected for use in all later test administrations. Care was exercised that the task example not reveal the nature of the cognitive skill which the test is to measure. Comments on the clarity and interest of all instructions were obtained from a panel of high school students (details on the composition of this panel is reported under STUDENT INTEREST) and revisions were made according to this feedback.

TEST VOCABULARY

The test battery is not intended as a measure of reading skills or of vocabulary development. To minimize confounding due to such factors, items were screened for words not available in an average ninth grader's vocabulary (Thorndike-Lorge, 1944). In addition Dale-Chall (1948) readability scores were computed for selected portions of the battery. Scores ranged from 7.5 to 8.2, indicating that test items are suitable for the average reading ability of Grades 7 and 8. These steps do not, of course, eliminate confounding due to differences in reading skills, but they should tend to minimize these differences insofar as possible.[1] In addition, they indicate the battery's appropriateness for the intended Grades (10–12).

STUDENT INTEREST

Immediately following Pretest One students were asked to rate the testimony portions of the battery on seven-step interest, readability, and difficulty scales. Testimony I was rated as quite readable and quite easy, while all

---

[1] The correlations of tests in this battery with various IQ and reading scores are available on request.

Table 1

Test Administrations in the Development of WISTTRA

	Tests administered	Place of testing	Date of testing	Sample size	Educational characteristics	Males	Females
Pretest One	TI, TII, TIII, RI	Madison, Wisc.	July, 1966	38	Students attending the Debate Program, Wisconsin High School Speech Institute; Grades 9, 10, and 11	23	15
Pretest Two	TI, TII, TIII, RI, RII, RIII RIV	Monona, Wisc.	Sept., 1966	58	Tenth, eleventh, and twelfth graders in an elective speech course at Monona Grove High School	27	31
Pretest Three	TI, TII, TIII, RI, RII, RIII, RIV	Lodi, Wisc.	Dec., 1966	187	Grades 7-12 of Lodi Junior and Senior High Schools	101	86
Pretest Four	TI, RI, RII, RIII, RIV	Juneau and Reeseville, Wisc.	Nov., 1967	258	Grades 10-12 of the Juneau and Reeseville High Schools	123	135
Normative Study	TI, TII, TIII, RI, RII, RIII, RIV	Clinton, Cedarburg, Reedsburg, and Owen-Withee, Wisc.	April, 1968	3090 to 3118[a]	Grades 7-12 of all four schools	1507 to 1515[a]	1583 to 1603[a]

[a]Variation due to student absenteeism during the testing period.

other ratings for the testimony tests were moderately readable, moderately easy, and moderately interesting. It should be remembered, however, that these students had received considerable instruction in argumentation and should have found the tests less challenging than students without special training in the area.

Item interest was also discussed with a panel of five high school sophomores who had no previous speech or argumentation course work. The panel was selected on the basis of Henman-Nelson IQ and SCAT Reading Scores to represent a range of abilities at that grade level. These are given in Table 2. Comments by the student consultants were considered during subsequent revisions of the test items.

Table 2

Panel of Student Consultants

Student	Henman-Nelson (IQ)	SCAT (Percentile)
A	121	99
B	121	94-99
C	143	99
D	107	70-83
E	112	85-96

SUBJECT MATTER OF ITEMS

Items were constructed using commonplace information from the subject-matter areas of

government, entertainment, and education. Fictional names of persons, places, and events were used where possible. Each test presents an approximate balance of items representing the three subject-matter areas. The data from Pretest One were examined for confounding due to subject area variables and revealed no significant differences in scores among the three areas. However, since items in these three areas deal with common topics of discussion, approximately equal representation of these subjects was retained.

CONTENT VALIDITY

As illustrated in Figure 1, WISTTRA was constructed to measure cognitive skills related to certain fundamental concepts of verbal argument. The three tests of testimony were designed to measure the student's ability to detect instances which violate common internal and external tests of testimony. The reasoning tests were designed to measure the student's ability to recognize the essential components of an argument, to ask relevant questions about arguments, and to draw correct conclusions from arguments.

Based upon pilot study information, subtests for Testimony I and Testimony III were retained as illustrated in Figure 1. The pilot study results indicated that subtests need not be retained for Testimony II and the four reasoning tests. Further study of the dimensionality of all the tests is in progress using factor analytic procedures.

At two points in the development of the tests —before Pretest One and prior to the Normative Study—the battery was submitted to panels of experts in the field of argumentation trained in the conceptual basis of the instrument. On both occasions three-judge panels were used. Following a Q-sort technique the judges were asked to place items in relevant categories or in a 'cannot tell' category. Criteria for categorizing items included (where relevant) argument type, type of rule violated, statement type, and completeness of argument. Judge agreement ranged from 94.9 to 98.9% for the tests coded in the initial stages of development and from 85.4 to 98.4% for the tests used in the normative study. The decline in coder

<table>
<tr><td colspan="8">CONCEPTS</td></tr>
<tr><td colspan="6">Appraising Testimony</td><td colspan="2">Appraising Reasoning</td></tr>
<tr><td colspan="5">Internal Tests</td><td colspan="2">External Tests</td><td>Data</td><td>Warrant</td><td colspan="2">Reservation</td><td>Claim</td></tr>
<tr><td colspan="5">TI
60 items</td><td>TII
20
items</td><td>TIII
40 items</td><td rowspan="3">Always given—no test</td><td>RI
28 items</td><td>RII
28
items</td><td>RIII
28
items</td><td>RIV
28 items</td></tr>
<tr><td></td><td colspan="4">Reject</td><td rowspan="3">Consistency</td><td rowspan="2">Recency</td><td rowspan="2">Proximity</td><td rowspan="2">Recognizing and selecting warrants</td><td rowspan="2">Recognizing statements which answer reservations</td><td rowspan="2">Selecting reservations</td><td rowspan="2">Selecting claims</td></tr>
<tr><td rowspan="2">Accept</td><td rowspan="2">Biased on topic</td><td rowspan="2">Not in position to observe</td><td rowspan="2">Incompetent to observe</td><td rowspan="2">Unqualified to judge</td><td>20
items</td><td>20
items</td></tr>
<tr><td>20
items</td><td>10
items</td><td>10
items</td><td>10
items</td><td>10
items</td><td></td><td></td><td colspan="5">Each of seven warrant types—sign, class, causal, alternative, parallel case, comparative, and warrant-supportive—is represented by four items in each reasoning test.</td></tr>
</table>

Figure 1. Relationship of WISTTRA to Concepts Identified

agreement is attributable to the fact that only items which achieved high coder agreement were used in drawing up the first edition of the tests while the pool of items coded on the second occasion consisted of all items comprising the normative edition of the tests.

INTERNAL CONSISTENCY RELIABILITY

Hoyt analysis of variance reliability estimates were obtained for all of the tests. This is an internal consistency measure of reliability and as such estimates consistency of performance on a relatively homogeneous power test.

Rigid standards for the interpretation of reliability estimates are not overly meaningful. As a rule of thumb, reliabilities of at least .80 are recommended for evaluating level of group accomplishment and .90 for evaluating level of individual accomplishment. In practice however, reliability estimates of .50 to .80 are often treated as indications of a relatively precise enough instrument for group differentiation. Thorndike and Hagen (1961) discuss this problem in terms of the percent of times the direction of difference will be reversed in subsequent testing for scores falling at the 75th and 50th percentiles for various values of reliability estimates. The security of a conclusion based upon a particular test increases much more rapidly for groups than it does for individuals as the reliability of the test increases. For example, the probability of reversal is one in three for scores of single individuals when the reliability is .50; the probability of reversal is 1 in 20 for means of groups of 25 when the reliability is .50.

The standard error of measurement is a second index of test consistency. This is a measure of the variability of the scores a subject would obtain on repeated measurements using the same test. The standard error of measurement indicates how much his obtained score for a single administration is likely to vary with repeated testing, i.e., how nearly "correct" this obtained score is. For a student's hypothetical distribution of repeated scores on a test, his obtained score would fall within one standard error value of his actual obtained score about two-thirds of the time.

Another way to look at the interpretation of a reliability estimate is in terms of the size of the standard error of measurement relative to the standard deviation of test scores. This is discussed by Thorndike (1951). If the reliability is zero the standard error of measurement would equal the standard deviation of the test. For reliability estimates of .80 and .90 the standard error of measurement is reduced to only 45% and 32% of its value for zero reliability.

Maximum reliabilities were sought for all tests but the researchers' expectations were conditioned by two considerations: (1) some of the tests are composed of subtests sufficiently divergent in character to reduce the overall homogeneity of the total tests (TI and TIII) and (2) some subtests are composed of so few items that high reliabilities are not likely (subtests of RI, RII, RIII, and RIV). For these reasoning tests the total test reliability should not have been affected by (2) except that item data for the subtests were used in selecting items for inclusion in the total test. The purpose of this action was to enable the researchers to obtain reliabilities for the four item subtests of a sufficient magnitude to enable further study of the dimensionality of the tests.

ITEM CHARACTERISTICS

During the development of the tests, items were continually revised to improve the instrument on the basis of item characteristic data obtained from the GITAP item analysis program (Baker 1966, 1968). This program provides difficulty level, biserial correlation, X_{50}, and β statistics for each choice of each item. In addition it gives descriptive statistics, the standard error of measurement, and the Hoyt reliability estimate for the total test. Certain item characteristic criteria were used in selecting and refining items on the basis of the GITAP results. Items to be retained in a revised edition of the test had to meet the minimum requirement as given for each of the following criteria for the correct choice:

1. Preferably fall within a middle difficulty range as defined by Ebel (1965). See Table 3.
2. Have a biserial correlation $\geq .30$.
3. Have an X_{50} between +2.00 and -2.00.
4. Have a $\beta \geq .30$.

In addition each incorrect choice had to meet the following minimum requirements:

1. Have a reasonable minimum proportion of subjects respond to it.
2. Have a biserial correlation $< -.25$ and preferably $< -.30$.
3. Have an X_{50} lower than the X_{50} for the correct choice.
4. Have a $\beta < -.25$ and preferably $< -.30$.

These criteria were established in consultation with staff of the R & D Center and on the basis of reasonably standard rules of thumb for item evaluation.

In a few cases where one or more choices of an item were slightly deficient in meeting one or more of the standards but it was felt that the item was still basically good, slight revisions were made in the item. In so far as possible it was desired that all items meet these criteria on the basis of each of two analyses—one with total test score as the criterion ability and another with appropriate subtest score as the criterion ability.

The difficulty of an item is indexed by the proportion of subjects who responded correctly to that item. Thus, the greater the value of the difficulty index the easier the item. An item of middle difficulty is defined by Ebel (1965) as one for which the proportion of correct responses is halfway between the expected chance proportion and 100%, and he further states that items in a midrange of difficulty —30% to 70% of the nonchance range—are almost as effective discriminators as are items of middle difficulty. This middle difficulty range was taken into consideration in defining desirable levels of difficulty for the items of WISTTRA. These levels are specified in Table 3. In general, in assembling the total test the items were ordered by increasing level of difficulty.

For the biserial correlation, X_{50}, and β item statistics, both the results for total test and appropriate subtest analyses were used. An attempt was made to use only items that met the standards on the basis of both analyses. In a few cases where this was impossible, the subtest analysis was the prime consideration.

The biserial correlation coefficient is an index of the discriminating ability of the item choice. For this analysis the criterion ability used was total test score. As with any correlation there is no rigid standard for interpreting a biserial correlation. Maximum correlations were desired for WISTTRA and .30 was set as a minimum for the correct choice. A low biserial correlation means that the item is not discriminating across the criterion ability range—a student who had a poor criterion score would be almost as likely to get the item correct as one who had a good criterion score. The negative biserial correlation for the incorrect choices indicates a descending slope of the regression line from left to right. Thus, poor students would be more likely to respond to those choices than would good students. The greater the absolute value of the correlation the greater the discriminating power of the item.

X_{50} is the point on the criterion scale, given in standard deviation units, corresponding to the median of the item characteristic curve and is the point at which the item choice has maximum discrimination. Figure 2 illustrates a typical item characteristic curve. Subjects with a criterion score equal to X_{50} have a 50-50 chance of choosing that response. Thus, +2.00 and -2.00 were used as desirable limits as this range would include approximately 99% of the cases. It was essential, for an item to be retained in the test, that the X_{50} value for all the incorrect choices be less than that for the correct choice with the exception of all two choice items. For two choice items the X_{50} value is the same for both choices.

β can be thought of as the slope of the item characteristic curve at the X_{50} point and is an index of the discrimination power of the item. The higher the β value the greater the slope of the curve and the more clearly the item is discriminating. The maximum positive βs were desired for the correct choice and negative ones required for all incorrect choices.

SUMMARY

In developing WISTTRA the researchers attempted to structure an instrument capable of

Table 3

Preferable Difficulty Levels for WISTTRA

Test	Number of Choices	Chance Probability	Middle Difficulty Point	Middle Difficulty Range
TI, TII TIII, RII	2	.500	.750	.650 to .850
RIV	3	.333	.667	.534 to .800
RIII	4	.250	.625	.475 to .775
RI	5	.200	.600	.440 to .760

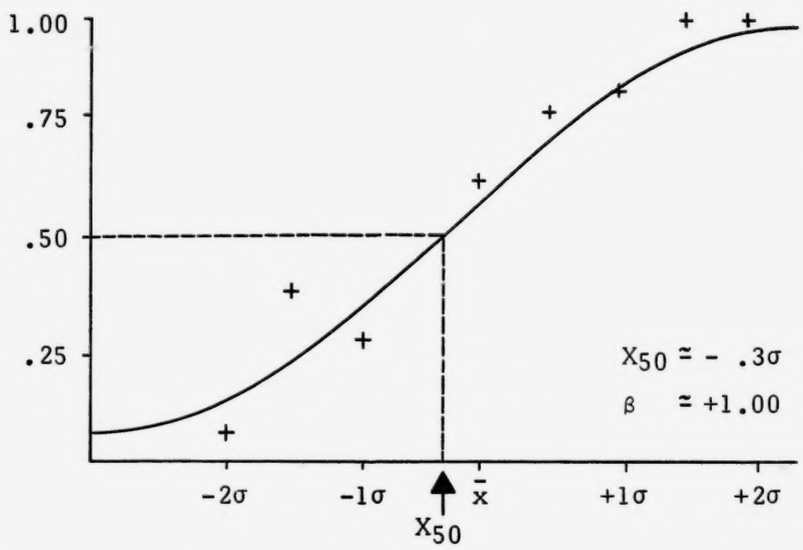

Figure 2. A Typical Item Characteristic Curve
Reproduced from Baker, 1965.

stable, precise discriminations across the broad ability ranges of high school students. At all stages standard procedures and criteria were employed to insure as much as possible that the final test battery would perform according to these expectations. Ideals of this sort must, however, be tempered by the demands of interestingness, practicability, and content validity, as well as by the fact that cognitive skills tend to elude precise measurement. This section has attempted to convey the criteria and considerations which the investigators employed to balance the often conflicting demands such a testing instrument must satisfy.

IV
DISCUSSION OF SPECIFIC TESTS

This section of the paper presents a discussion of WISTTRA on a test by test basis providing a statement of each test's objectives and a brief description of each test as it exists in its latest edition.

TESTIMONY I: Appraising Testimony in Terms of Internal Criteria

Objective

Testimony I is designed to measure a student's ability to use the internal tests of testimony to discriminate between reliable and unreliable instances of testimony.

Structure of the Test

Testimony I consists of 60 two-choice items. Each item presents the name of a source (person, office, publication, etc.) and a statement (generalization, quality judgment, statistic, etc.) made by that source on some particular topic. Consider the following two examples:

1. Senate Reporter: Sixty-five percent of the Republican Senators voted for the Smith-Doe Bill.
2. High School Student: Sun Village is an excellent example of modern literature.

Students are asked to indicate whether they would accept or reject that statement on the grounds that it was made by that source. Four criteria for acceptance are represented in the items—(1) Is the source in a position to observe? (2) Is the source competent to observe? (3) Is the source unbiased? and (4) Is the source qualified to judge? Positive (accept) items, such as Example 1 above, are those which meet all four criteria; negative (reject) items meet three of the four, but do not fulfill the fourth criterion (e.g., in position, competent, unbiased, but not qualified to judge), as in Example 2 above. A student responding correctly on all items would accept 20 instances and reject 40. Violations of internal criteria are distributed across reject items in a balanced fashion so that each criterion is represented by 10 reject items.

There are indications that the five subtests of Testimony I should be kept as individual tests and not grouped into one composite test. Further study on the dimensionality of the tests is in progress and will be reported in A Factor Analytic Study of the Wisconsin Tests of Testimony and Reasoning Assessment (Harris, 1969).

TESTIMONY II: Appraising Testimony in Terms of External Criteria: Consistency with Other Testimony

Objective

Testimony II is designed to measure a student's ability to recognize inconsistency between two instances of testimony.

Structure of the Test

Testimony II consists of 20 two-choice items. Each item presents two similar statements attributed to the same source or to different sources. Two examples are:

1A. Sam, Pro Golf Official: The greens are in good condition for today's match.

B. Sam, Pro Golf Official: Even the best of the golfers are complaining about the rough spots on the greens in today's match.

2A. American League President: Public interest in baseball has declined in the last year.

B. National League President: All major league clubs have had sizable increases in attendance over the past year.

Students are asked to compare the two instances of testimony and determine whether accepting the second instance would make them more or less likely to accept the first. In the two examples above, the second instance clearly makes the first instance less likely. These items represent the single criterion of consistency between instances of testimony using statements by the same source or statements by different sources. Testimony II presents 10 consistent and 10 inconsistent pairs of testimony. Both consistent and inconsistent items are balanced so that single-source pairs and two-source pairs appear with the same frequency.

TESTIMONY III: Appraising Testimony in Terms of External Criteria: Recency and Proximity

Objective

Testimony III is designed to measure a student's ability to use the external tests of proximity and recency to discriminate between reliable and unreliable instances of testimony.

Structure of the Test

Testimony III consists of 40 two-choice items. Each item presents a pair of similar statements attributed to different sources. Proximity differences (primary versus secondary information) are contrasted within 20 of the 40 items by structuring one member of the pair as a reference to some other source and the second member of the pair as a source of the sort referred to by the first. For example:

1A. Health Inspector: The pool's chlorine content was too high today when I tested it.

B. Pool Director: I was told that today's tests indicated the concentration of chlorine in the water to be above the suggested limit.

In this case, the health inspector's statement is to be preferred since he is a primary source reporting a direct observation whereas the pool director is a secondary source who is merely reporting what someone else told him. Recency differences (recent vs. out-of-date information) are contrasted in the remaining 20 items by including time references which indicate that one statement of the pair is more recent than the other. For example:

2A. Teacher A: As we begin the 1966–67 school year we can expect more than 75 books to disappear from the library.

B. Teacher B: Less than fifty books were missing from the library when the 1966–67 school year ended.

In this case, when the subject in question is the number of books disappearing during a school year, the testimony of Teacher B is to be preferred since it is the more recent. In order to measure a student's ability to handle these criteria in realistic settings, each of the subtests presents 10 pairs of consistent instances of testimony and 10 pairs of inconsistent instances. The student is asked to select the best instance of testimony independent of any inclination to agree with either of the sources.

Previous preliminary study indicates that Testimony III consists of two somewhat independent subtests. Reliability and item characteristic information will be given for Testimony III as a composite and as two separate subtests. Further study on the dimensionality of the tests is in progress.

REASONING I: Recognizing and Selecting Warrants in Arguments

Objective

Reasoning I is designed to measure a student's ability to recognize the absence of warrants and to select appropriate warrants when needed.

Structure of the Test

Reasoning I consists of 28 five-option multiple choice items. Each item presents an argument based on reasoning which is either complete (data, warrant, and claim) or incomplete (data and claim, but irrelevant information instead of warrant). Four possible warrants and a none-needed option are given for each item. For example:

Mr. Henswar, who moved into our neighborhood last week, is a judge. He is one of

the new judges I have never seen. It can be concluded that Mr. Henswar is a dignified man.

A. Mr. Henswar is a typical newly appointed judge.
B. Repeated presence in court is a sign of dignity in a man.
C. Judges are dignified men.
D. Mr. Henswar is more dignified than our previous judge.
E. None needed.

Students are asked to mark option E if the argument is complete as it stands in the initial paragraph. If not, they are to select the appropriate warrant from options A–D. For instance, since the warrant is absent in the above example, the student should select response C which provides an appropriate inference license for a class argument. A student responding correctly to all questions will select E for 10 of the 28 items. Each item represents one of seven argument types—sign, cause, class, comparative, parallel case, alternative, and warrant-supportive. Each of the seven argument types is represented by four items, two or three of which require completion by appropriate warrant selection. In addition, an effort was made to distribute warrant types among the distractor options in a balanced fashion. Thus, in the above example, the incorrect responses A, B, and D are warrants for warrant-supportive, sign, and comparative arguments respectively.

REASONING II: Recognizing Statements Which Answer Reservations in Arguments

Objective

Reasoning II is designed to measure the student's recognition of statements in arguments which anticipate and answer reservations.

Structure of the Test

Reasoning II consists of 28 two-choice items. Each item presents a pair of complete arguments containing data, warrant, claim, and some additional information. The student is instructed to indicate which of the two arguments is better. For example:

A. George is probably a good young farmer because he is a member of our school's Future Farmers of America Club. Last year he won four blue ribbons at the county fair with his dairy cattle. Our county agricultural agent says that membership in the FFA is a pretty good sign that a boy is a good young farmer.

B. George is probably a good young farmer because he is a member of our school's Future Farmers of America Club. The Club has twenty-two members and meets every Saturday morning in the Agricultural Lab. Our county agricultural agent says that membership in the FFA is a pretty good sign that a boy is a good young farmer.

The paired arguments are alike except that in one of them the additional information is irrelevant to the argument while in the other the information removes or answers a possible refutation (reservation) of the argument. Thus, in the above example, the first argument is to be preferred to the second because the first contains information which answers the "lack of concurrent sign" reservation. All seven argument types are represented equally in the items. Reasoning II contains four items for each of the seven argument types: sign, cause, class, comparative, parallel case, alternative, and warrant-supportive.

REASONING III: Selecting Reservations in Arguments

Objective

Reasoning III is designed to measure the student's ability to discriminate between relevant and irrelevant reservations.

Structure of the Test

Reasoning III consists of 28 four-option multiple choice items. Each item presents a complete argument (data, warrant, and claim) and four statements to be considered as possible reservations to the argument. For example:

Enrollment in our high school has been steadily increasing. Since increases in enrollment force school boards to build new high schools, our school board will probably build another high school soon.

A. Unless there is still plenty of room in the old high school.
B. Unless all school boards face increasing enrollments.
C. Unless our school system has fewer students than many other systems which built new high schools.

D. Unless the relationship between enrollments expressed by the word "increased" does not imply future changes in enrollment.

The student is asked to select from the four the reservation which best qualifies or refutes the argument. In each case only one of the four choices is a relevant reservation to the argument type represented. The other three responses, although appearing as reservations in terms of phrasing, do not lessen the confidence which may be placed in the claim advanced. Thus, answer A in the example above represents the partial cause reservation to a cause-effect argument and is therefore the appropriate response while the other three responses simply provide seemingly relevant information in a reservation form inappropriate to a causal argument.

Reasoning III presents four items for each of the seven argument types. This format enabled the investigators to use each of the various reservation types at least once in connection with an appropriate argument type and to roughly balance the distribution of reservation types across distractor choices.

REASONING IV: Selecting Claims in Arguments

Objective

Reasoning IV is designed to measure a student's ability to select the claim appropriate to a given argument.

Structure of the Test

Reasoning IV consists of 28 three-option multiple choice items. Each item presents the data, warrant, and some additional information for an argument. The items are systematically varied such that the additional information is irrelevant in 9 items, provides an answer to a reservation in 10 other items, and raises a reservation (with no answer given) in the remaining 9 items. The student is instructed to select the proper claim to the argument (of two choices given) or to indicate that it is not possible to make a proper claim given the information presented in the argument. For example:

> Johnny always turns his work in on time. Turning work in on time plays an important part in passing college courses. Sometimes Johnny spends little time on his work and does not care much about studying. Therefore:
>
> A. Johnny probably will pass his college courses.
> B. You really can't tell whether Johnny will pass his college courses.
> C. Johnny probably does careful work in his courses.

The three possible answers are constructed such that one states a topically related idea which does not follow the structure of the data and warrant, one denies that a particular conclusion is possible, and one presents a straightforward claim. A student responding correctly would select the "cannot tell" option for items with unanswered reservations (as in the example above) and the straightforward claim in all other cases. Again each of the argument types is represented by four items.

V

RELIABILITY ESTIMATES AND ITEM STATISTICS

The reliability estimates and item statistics reported in this section were obtained in the normative study. In all cases, they are given separately for each sex group for each grade (seven through twelve).

SAMPLE

The total number of subjects tested ranged from 3090 to 3118 for any one test. The total number of subjects within a single age and sex group ranged from 190 to 311 for any one test. These subjects were obtained by randomly sampling schools from a single stratification of the population of Wisconsin school districts. This was accomplished by using the results of a study by Miller et al. (1967) which describes Wisconsin school districts on the basis of factor scores for a number of factors. The following five factors were used in identifying a homogeneous stratified population for the study: (1) numerical size, (2) organizational complexity, (3) teacher experience, (4) economic power, and (5) size of school unit. For further details on the population and sampling procedures used refer to A Study of Student Abilities in the Evaluation of Verbal Argument (Rott, Feezel, & Allen, in press).

RELIABILITY ESTIMATES

Reliability estimates were computed using the Hoyt analysis of variance procedures and were obtained as part of the results of the Generalized Item and Test Analysis Program (Baker, 1968) used to analyze the tests. These estimates are presented in Tables 4 through 17 for each of the seven tests as a total test and for subtests of Testimony I and Testimony III. Also included in these tables, for each grade and sex group, is the sample size, mean, standard deviation, range, and standard error of measurement.

ITEM STATISTICS

A summary of the item statistics (difficulty, biserial correlation, X_{50}, and β) for the correct choices for each of the seven tests as a total test and for subtests of Testimony I and Testimony III are given in Tables 18 through 31. The investigators realize there are problems with using the mean as a measure of central tendency for the biserial correlation and β since they are not linear, but it was felt the mean would give the reader some indication of central tendency and at least show the general increase in the value of these statistics from grade seven through grade twelve.

Table 4

Testimony I: Reliability Estimates for the Total Test

Grade/Sex	N	Mean Score (60 max.)	Standard Deviation	Range of Scores	Standard Error of Measurement	Hoyt Reliability
7M	246	37.89	7.44	18–58	3.48	.78
7F	251	38.96	6.62	12–56	3.42	.73
8M	228	39.78	7.85	24–54	3.38	.81
8F	224	40.87	7.06	25–55	3.29	.78
9M	304	41.00	8.04	13–56	3.29	.83
9F	302	42.22	7.23	25–57	3.19	.80
10M	287	43.69	8.04	6–58	3.08	.85
10F	302	44.53	6.70	23–57	2.97	.80
11M	253	45.45	6.84	23–57	2.92	.81
11F	265	44.38	6.45	16–56	2.96	.79
12M	190	44.98	8.09	22–59	2.97	.86
12F	253	45.30	6.73	12–57	2.88	.81

Table 5

Testimony I: Reliability Estimates for the Accept Subtest

Grade/Sex	N	Mean Score (20 max.)	Standard Deviation	Range of Scores	Standard Error or Measurement	Hoyt Reliability
7M	246	14.22	3.90	0–20	1.80	.78
7F	251	15.00	3.23	5–20	1.75	.69
8M	228	15.00	3.50	7–20	1.74	.74
8F	224	15.65	3.16	7–20	1.64	.72
9M	304	15.42	3.57	6–20	1.66	.77
9F	302	16.19	3.05	5–20	1.54	.73
10M	287	16.43	3.46	7–20	1.47	.80
10F	302	17.14	2.48	8–20	1.37	.68
11M	253	17.50	2.61	7–20	1.31	.74
11F	265	17.38	2.53	6–20	1.31	.72
12M	190	17.14	3.21	7–20	1.37	.81
12F	253	17.68	2.41	4–20	1.23	.72

Table 6

Testimony I: Reliability Estimates for the Bias Subtest

Grade/Sex	N	Mean Score (10 max.)	Standard Deviation	Range of Scores	Standard Error of Measurement	Hoyt Reliability
7M	246	5.16	2.17	0–9	1.32	.59
7F	251	5.19	2.14	1–10	1.29	.60
8M	228	5.48	2.05	0–10	1.34	.53
8F	224	5.38	2.28	0–10	1.26	.66
9M	304	5.53	2.15	0–10	1.33	.57
9F	302	5.49	2.12	0–10	1.27	.60
10M	287	5.86	2.13	0–10	1.27	.61
10F	302	5.57	2.20	1–10	1.21	.66
11M	253	6.00	2.21	1–10	1.21	.67
11F	265	5.62	2.19	0–10	1.20	.67
12M	190	5.91	2.15	0–10	1.26	.61
12F	253	5.74	2.31	0–10	1.19	.70

Table 7

Testimony I: Reliability Estimates for the Position Subtest

Grade/Sex	N	Mean Score (10 max.)	Standard Deviation	Range of Scores	Standard Error of Measurement	Hoyt Reliability
7M	246	5.65	1.91	0–10	1.41	.39
7F	251	5.59	2.00	1–10	1.40	.46
8M	228	5.91	1.92	0–10	1.40	.41
8F	224	5.91	1.90	1–10	1.37	.43
9M	304	6.19	1.88	2–10	1.36	.42
9F	302	6.36	2.09	0–10	1.32	.56
10M	287	6.50	1.96	1–10	1.30	.51
10F	302	6.66	1.90	1–10	1.27	.50
11M	253	6.47	1.92	1–10	1.29	.50
11F	265	6.42	1.93	2–10	1.28	.51
12M	190	6.54	2.08	1–10	1.26	.59
12F	253	6.67	2.03	1–10	1.22	.60

Table 8

Testimony I: Reliability Estimates for the Competence Subtest

Grade/Sex	N	Mean Score (10 max.)	Standard Deviation	Range of Scores	Standard Error of Measurement	Hoyt Reliability
7M	246	6.55	2.03	1–10	1.34	.52
7F	251	6.34	2.11	1–10	1.35	.55
8M	228	6.75	2.01	2–10	1.32	.52
8F	224	6.88	2.07	1–10	1.29	.57
9M	304	6.94	2.04	2–10	1.28	.57
9F	302	6.88	2.18	1–10	1.26	.63
10M	287	7.45	2.01	2–10	1.20	.60
10F	302	7.30	2.13	2–10	1.19	.65
11M	253	7.63	1.89	2–10	1.16	.58
11F	265	7.09	2.10	2–10	1.22	.62
12M	190	7.55	1.96	1–10	1.18	.60
12F	253	7.32	1.97	2–10	1.20	.58

Table 9

Testimony I: Reliability Estimates for the Qualification Subtest

Grade/Sex	N	Mean Score (10 max.)	Standard Deviation	Range of Scores	Standard Error of Measurement	Hoyt Reliability
7M	246	6.31	2.01	0–10	1.37	.48
7F	251	6.84	1.75	2–10	1.33	.36
8M	228	6.65	2.02	2–10	1.32	.52
8F	224	7.05	1.97	3–10	1.28	.53
9M	304	6.91	1.98	2–10	1.28	.53
9F	302	7.30	1.90	2–10	1.24	.53
10M	287	7.45	1.92	2–10	1.21	.56
10F	302	7.88	1.73	3–10	1.12	.53
11M	253	7.86	1.67	3–10	1.14	.48
11F	265	7.89	1.72	3–10	1.13	.52
12M	190	7.85	1.91	3–10	1.12	.61
12F	253	7.88	1.72	2–10	1.12	.53

Table 10

Testimony II: Reliability Estimates

Grade/Sex	N	Mean Score (20 max.)	Standard Deviation	Range of Scores	Standard Error of Measurement	Hoyt Reliability
7M	246	13.14	3.83	0–20	1.93	.73
7F	251	12.77	3.86	4–20	1.95	.73
8M	228	12.67	4.32	4–20	1.92	.79
8F	224	13.16	4.26	6–20	1.88	.79
9M	304	12.75	4.36	4–20	1.91	.80
9F	302	13.26	4.51	4–20	1.84	.82
10M	287	13.33	4.89	2–20	1.80	.86
10F	302	15.14	4.32	5–20	1.64	.85
11M	253	14.60	4.58	0–20	1.69	.86
11F	265	14.83	4.79	3–20	1.63	.88
12M	190	14.52	4.76	6–20	1.68	.87
12F	253	14.98	4.69	0–20	1.61	.88

Table 11

Testimony III: Reliability Estimates for the Total Test

Grade/Sex	N	Mean Score (40 max.)	Standard Deviation	Range of Scores	Standard Error of Measurement	Hoyt Reliability
7M	246	23.59	5.58	12–37	2.93	.72
7F	251	23.55	5.33	14–37	2.90	.70
8M	227	24.04	6.14	5–40	2.89	.77
8F	223	25.25	5.29	14–38	2.80	.71
9M	303	24.40	5.99	11–38	2.86	.77
9F	305	25.68	5.72	8–38	2.77	.76
10M	228	26.28	6.54	3–40	2.73	.82
10F	311	27.05	5.83	16–39	2.65	.79
11M	256	27.33	6.66	0–40	2.64	.84
11F	262	28.36	6.07	10–39	2.55	.82
12M	195	26.82	6.42	13–39	2.70	.82
12F	251	27.69	6.33	4–40	2.56	.83

Table 12

Testimony III: Reliability Estimates for the Recency Subtest

Grade/Sex	N	Mean Score (20 max.)	Standard Deviation	Range of Scores	Standard Error of Measurement	Hoyt Reliability
7M	246	12.57	3.32	4–20	2.00	.62
7F	251	12.88	3.42	7–20	1.97	.65
8M	227	12.73	3.61	6–20	1.97	.69
8F	223	13.73	3.44	7–20	1.86	.69
9M	303	12.91	3.64	5–20	1.95	.70
9F	305	13.90	3.36	7–20	1.85	.68
10M	288	14.15	3.64	7–20	1.83	.73
10F	311	14.69	3.21	6–20	1.75	.71
11M	256	14.54	3.70	0–20	1.76	.76
11F	262	15.40	3.28	7–20	1.66	.73
12M	195	14.30	3.64	6–20	1.79	.75
12F	251	14.97	3.25	7–20	1.69	.72

Table 13

Testimony III: Reliability Estimates for the Proximity Subtest

Grade/Sex	N	Mean Score (20 max.)	Standard Deviation	Range of Scores	Standard Error of Measurement	Hoyt Reliability
7M	246	11.02	3.22	3–20	2.06	.57
7F	251	10.67	3.07	4–19	2.05	.53
8M	227	11.31	3.53	5–20	2.03	.65
8F	223	11.52	3.09	4–19	2.01	.56
9M	303	11.49	3.40	4–20	2.02	.63
9F	305	11.78	3.42	2–19	1.98	.65
10M	288	12.13	3.87	1–20	1.94	.73
10F	311	12.63	3.71	2–20	1.90	.72
11M	256	12.79	3.86	0–20	1.89	.74
11F	262	12.96	3.81	4–20	1.84	.75
12M	195	12.52	3.64	4–20	1.94	.70
12F	251	12.72	3.99	4–20	1.83	.78

Table 14

Reasoning I: Reliability Estimates

Grade/Sex	N	Mean Score (28 max.)	Standard Deviation	Range of Scores	Standard Error of Measurement	Hoyt Reliability
7M	245	8.24	5.00	0–24	2.18	.80
7F	248	8.97	4.87	0–25	2.24	.78
8M	230	9.41	5.58	1–28	2.23	.83
8F	218	10.79	5.99	3–27	2.24	.85
9M	302	9.69	5.77	1–25	2.23	.84
9F	301	11.61	6.57	1–27	2.24	.88
10M	277	12.90	6.89	1–28	2.25	.89
10F	294	14.48	7.48	2–28	2.16	.91
11M	262	14.08	7.58	1–28	2.18	.91
11F	270	15.21	7.48	1–28	2.16	.91
12M	191	13.86	7.84	1–28	2.16	.92
12F	252	15.70	7.16	3–28	2.18	.90

Table 15

Reasoning II: Reliability Estimates

Grade/Sex	N	Mean Score (28 max.)	Standard Deviation	Range of Scores	Standard Error of Measurement	Hoyt Reliability
7M	246	15.94	3.91	7–26	2.42	.60
7F	251	17.24	4.16	9–27	2.37	.66
8M	227	17.76	4.40	8–28	2.38	.70
8F	223	18.25	4.43	9–27	2.30	.72
9M	303	18.00	4.87	8–28	2.32	.76
9F	305	19.55	4.62	7–28	2.17	.77
10M	288	18.88	4.95	6–27	2.14	.81
10F	311	21.00	4.85	9–28	2.00	.82
11M	256	20.81	5.19	0–28	2.05	.84
11F	262	21.50	4.53	9–28	1.94	.81
12M	195	21.02	5.15	9–28	2.03	.84
12F	251	21.90	4.67	5–28	1.86	.84

Table 16

Reasoning III: Reliability Estimates

Grade/Sex	N	Mean Score (28 max.)	Standard Deviation	Range of Scores	Standard Error of Measurement	Hoyt Reliability
7M	245	12.44	5.67	3–27	2.34	.82
7F	248	13.77	5.24	3–26	2.35	.80
8M	230	13.89	6.13	3–28	2.31	.85
8F	218	15.72	5.89	4–26	2.27	.85
9M	302	14.66	6.36	3–26	2.27	.87
9F	301	16.57	5.92	0–27	2.24	.85
10M	277	17.55	6.22	3–28	2.18	.87
10F	294	19.14	5.42	5–28	2.10	.84
11M	262	18.68	6.25	0–28	2.11	.88
11F	270	19.53	5.63	2–28	2.06	.86
12M	191	18.03	6.71	4–28	2.12	.90
12F	252	20.59	4.94	5–28	2.00	.83

Table 17

Reasoning IV: Reliability Estimates

Grade/Sex	N	Mean Score (28 max.)	Standard Deviation	Range of Scores	Standard Error of Measurement	Hoyt Reliability
7M	245	14.39	4.97	4–27	2.40	.76
7F	248	15.41	4.73	5–27	2.39	.74
8M	230	16.06	4.81	5–25	2.37	.75
8F	218	16.94	4.86	5–27	2.33	.76
9M	302	16.60	5.25	4–27	2.30	.80
9F	301	17.82	5.09	4–28	2.25	.80
10M	277	18.72	5.12	5–28	2.21	.81
10F	294	19.73	4.51	6–28	2.15	.77
11M	262	19.23	4.96	3–28	2.17	.80
11F	270	20.17	4.35	6–28	2.10	.76
12M	191	19.04	5.63	5–28	2.14	.85
12F	252	21.16	3.88	10–28	2.03	.72

Table 18

Testimony I: Item Statistics for the Total Test
(60 Items)

Grade/Sex	N	Number of Items Within Middle Difficulty Range (.65—.85)	Range of Item Difficulties	Mean Difficulty	Number of Items with Biserial Correlations ≥ .30	Range of Biserial Correlations	Mean Biserial Correlation	Number of Low X_{50}s (-1.01 or below)	Number of Medium X_{50}s (-1.00 to +1.00)	Number of High X_{50}s (+1.01 or above)	Range of βs	Mean β
7M	246	32	.28-.82	.62	37	-.11 to .60	.36	30	24	6	-.10 to .75	.40
7F	251	34	.12-.89	.63	38	-.06 to .65	.31	36	18	6	-.06 to .86	.37
8M	228	33	.31-.86	.66	43	-.07 to .72	.39	37	21	2	-.07 to 1.04	.45
8F	224	30	.30-.89	.66	41	.05 to .67	.38	41	12	7	.05 to .91	.43
9M	304	38	.34-.86	.64	47	-.02 to .73	.42	41	15	4	-.02 to 1.06	.50
9F	302	32	.21-.94	.70	46	-.03 to .70	.40	43	14	3	-.03 to .98	.46
10M	287	38	.32-.92	.72	48	.00 to .80	.48	48	7	5	.00 to 1.32	.61
10F	302	16	.27-.97	.72	46	.05 to .80	.44	48	7	5	.05 to 1.31	.53
11M	253	26	.29-.94	.73	50	.08 to .83	.46	50	6	4	.08 to 1.49	.56
11F	265	22	.19-.95	.72	47	.09 to .67	.41	46	9	5	.09 to .91	.47
12M	190	30	.28-.93	.73	51	.07 to .90	.51	49	6	5	.07 to 2.09	.74
12F	253	22	.21-.97	.75	51	.02 to .89	.46	44	12	4	.02 to 1.96	.57

GPO 812—508—4

Table 19

Testimony I: Item Statistics for the Accept Subtest
(20 Items)

Grade/Sex	N	Number of Items Within Middle Difficulty Range (.65—.85)	Range of Item Difficulties	Mean Difficulty	Number of Items With Biserial Correlations ≥ .30	Range of Biserial Correlations	Mean Biserial Correlation	Number of Low X_{50}s (-1.01 or below)	Number of Medium X_{50}s (-1.00 to +1.00)	Number of High X_{50}s (+1.01 or above)	Range of βs	Mean β
7M	246	17	.44-.82	.71	20	.43 to .82	.59	10	10	0	.48 to 1.12	.76
7F	251	17	.45-.89	.75	19	.20 to .71	.53	15	5	0	.21 to 1.03	.66
8M	228	18	.50-.86	.75	20	.43 to .74	.57	15	5	0	.48 to 1.12	.71
8F	224	12	.45-.91	.78	20	.35 to .82	.58	18	2	0	.37 to 1.43	.76
9M	304	16	.56-.86	.77	20	.32 to .81	.63	16	4	0	.33 to 1.42	.85
9F	302	11	.49-.94	.81	20	.42 to .84	.62	17	3	0	.46 to 1.57	.85
10M	287	9	.58-.92	.82	20	.44 to .92	.72	18	2	0	.49 to 2.47	1.18
10F	302	3	.51-.97	.86	20	.36 to .86	.66	18	2	0	.42 to 1.71	1.20
11M	253	4	.59-.94	.88	20	.49 to 1.00	.71	19	1	0	.56 to 88.63	1.02[c]
11F	265	2	.54-.95	.87	20	.44 to 1.17	.71	18	1[a]	0	.49 to 2.50[b]	1.14[d]
12M	190	6	.68-.93	.86	20	.38 to 1.06	.76	18	1[a]	0	.42 to 9.70[b]	1.32[d]
12F	253	2	.55-.97	.88	20	.53 to 1.01	.76	15	3[a]	0	.62 to 3.27[b]	1.27[e]

[a]The X_{50} could not be computed for the items for which the biserial correlation exceeded 1.00.

[b]The highest β could not be computed since the highest biserial correlation exceeded 1.00.

[c]Based on 19 items excluding the β of 88.63.

[d]Based on 19 items.

[e]Based on 18 items.

Table 20

Testimony I: Item Statistics for the Bias Subtest
(10 Items)

Grade/Sex	N	Number of Items Within Middle Difficulty Range (.65—.85)	Range of Item Difficulties	Mean Difficulty	Number of Items with Biserial Correlations ≥ .30	Range of Biserial Correlations	Mean Biserial Correlation	Number of Low X_{50}s (-1.01 or below)	Number of Medium X_{50}s (-1.00 to +1.00)	Number of High X_{50}s (+1.01 or above)	Range of βs	Mean β
7M	246	4	.28-.74	.52	10	.47 to .77	.60	1	7	2	.52 to 1.22	.78
7F	251	3	.32-.77	.52	10	.44 to .75	.61	2	6	2	.49 to 1.14	.79
8M	228	4	.31-.73	.55	10	.39 to .68	.58	1	9	0	.42 to .93	.71
8F	224	4	.22-.77	.54	10	.52 to .75	.65	1	8	1	.61 to 1.14	.89
9M	304	4	.34-.73	.55	10	.43 to .70	.59	0	10	0	.48 to 1.00	.76
9F	302	5	.21-.76	.50	10	.46 to .75	.62	2	7	1	.52 to 1.15	.81
10M	287	6	.33-.77	.58	10	.52 to .75	.63	3	7	0	.61 to 1.14	.83
10F	302	5	.22-.79	.60	10	.50 to .80	.67	2	7	1	.58 to 1.45	.98
11M	253	6	.29-.82	.60	10	.49 to .82	.68	3	7	0	.57 to 1.43	.98
11F	265	4	.19-.81	.56	10	.60 to .83	.69	2	7	1	.64 to 1.52	.99
12M	190	6	.31-.80	.59	10	.45 to .80	.64	3	7	0	.51 to 1.37	.88
12F	253	6	.21-.83	.57	10	.60 to .83	.71	4	6	0	.66 to 1.49	1.04

Table 21

Testimony I: Item Statistics for the Position Subtest
(10 Items)

Grade/Sex	N	Number of Items Within Middle Difficulty Range (.65—.85)	Range of Item Difficulties	Mean Difficulty	Number of Items With Biserial Correlations ≥ .30	Range of Biserial Correlations	Mean Biserial Correlation	Number of Low X_{50}s (-1.01 or below)	Number of Medium X_{50}s (-1.00 to +1.00)	Number of High X_{50}s (+1.01 or above)	Range of βs	Mean β
7M	246	1	.31-.70	.56	10	.32 to .64	.50	4	6	0	.34 to .84	.59
7F	251	2	.34-.71	.56	10	.42 to .62	.53	0	10	0	.46 to .79	.63
8M	228	2	.35-.71	.59	10	.30 to .59	.51	1	8	1	.30 to .71	.61
8F	224	2	.32-.75	.59	10	.40 to .74	.53	1	8	1	.43 to 1.11	.65
9M	304	4	.38-.94	.62	10	.35 to .68	.53	2	7	1	.38 to .94	.63
9F	302	5	.34-.80	.64	10	.45 to .70	.59	4	6	0	.51 to 1.00	.75
10M	287	7	.32-.83	.65	10	.43 to .72	.58	5	5	0	.48 to 1.04	.73
10F	302	6	.29-.86	.67	10	.48 to .68	.58	4	5	1	.56 to .93	.72
11M	253	5	.31-.84	.65	10	.46 to .73	.58	4	5	1	.54 to 1.08	.73
11F	265	5	.28-.86	.64	10	.48 to .64	.58	3	7	0	.56 to .84	.71
12M	190	5	.28-.87	.66	10	.45 to .88	.63	2	8	0	.51 to 1.85	.88
12F	253	5	.25-.90	.67	10	.53 to .69	.64	3	7	0	.63 to .96	.84

Table 22

Testimony I: Item Statistics for the Competence Subtest
(10 Items)

Grade/Sex	N	Number of Items Within Middle Difficulty Range (.65—.85)	Range of Item Difficulties	Mean Difficulty	Number of Items With Biserial Correlations ≥ .30	Range of Biserial Correlations	Mean Biserial Correlation	Number of Low X_{50}s (-1.01 or below)	Number of Medium X_{50}s (-1.00 to +1.00)	Number of High X_{50}s (+1.01 or above)	Range of βs	Mean β
7M	246	5	.52-.73	.66	10	.48 to .68	.56	3	7	0	.52 to .95	.69
7F	251	5	.51-.73	.64	10	.41 to .68	.57	2	8	0	.45 to .92	.68
8M	228	6	.56-.78	.68	10	.33 to .68	.57	3	7	0	.35 to .93	.71
8F	224	5	.56-.81	.69	10	.45 to .75	.61	5	5	0	.51 to 1.14	.78
9M	304	6	.55-.80	.70	10	.46 to .67	.60	4	6	0	.53 to .91	.68
9F	302	5	.59-.71	.69	10	.45 to .77	.64	0	10	0	.58 to 1.59	.88
10M	287	8	.60-.86	.74	10	.47 to .79	.65	5	5	0	.54 to 1.32	.91
10F	302	6	.59-.88	.73	10	.48 to .86	.68	4	6	0	.54 to 1.72	1.02
11M	253	5	.61-.93	.76	10	.43 to .82	.64	4	6	0	.47 to 1.48	.90
11F	265	5	.53-.87	.71	10	.34 to .76	.64	4	6	0	.36 to 1.17	.89
12M	190	6	.61-.90	.76	10	.52 to .74	.66	5	5	0	.61 to 1.11	.89
12F	253	4	.56-.91	.73	10	.31 to .81	.64	4	6	0	.31 to 1.41	.89

Table 23

Testimony I: Item Statistics for the Qualification Subtest
(10 Items)

Grade/Sex	N	Number of Items Within Middle Difficulty Range (.65—.85)	Range of Item Difficulties	Mean Difficulty	Number of Items with Biserial Correlations ≥ .30	Range of Biserial Correlations	Mean Biserial Correlation	Number of Low X_{50}s (-1.01 or below)	Number of Medium X_{50}s (-1.00 to +1.00)	Number of High X_{50}s (+1.01 or above)	Range of βs	Mean β
7M	246	5	.45-.76	.63	9	.28 to .66	.55	1	9	0	.29 to .89	.67
7F	251	7	.50-.79	.69	9	.24 to .72	.52	6	4	0	.25 to 1.04	.63
8M	228	4	.51-.81	.66	10	.37 to .69	.58	3	7	0	.40 to .96	.72
8F	224	7	.56-.82	.71	10	.31 to .73	.59	5	5	0	.33 to 1.09	.76
9M	304	7	.48-.83	.69	10	.32 to .84	.60	5	5	0	.34 to 1.59	.80
9F	302	6	.54-.87	.73	10	.34 to .70	.60	5	5	0	.36 to .96	.79
10M	287	9	.58-.88	.75	10	.28 to .80	.64	6	4	0	.29 to 1.39	.90
10F	302	5	.55-.90	.79	10	.34 to .84	.67	8	2	0	.37 to 1.61	.99
11M	253	6	.53-.90	.79	10	.41 to .69	.62	7	3	0	.45 to .97	.81
11F	265	5	.56-.90	.79	10	.47 to .74	.65	6	4	0	.54 to 1.15	.88
12M	190	7	.58-.88	.78	10	.34 to .84	.69	8	2	0	.36 to 1.86	1.08
12F	253	5	.54-.92	.79	10	.37 to .86	.65	7	3	0	.40 to 1.68	.94

Table 24

Testimony II: Item Statistics
(20 Items)

Grade/Sex	N	Number of Items Within Middle Difficulty Range (.65—.85)	Range of Item Difficulties	Mean Difficulty	Number of Items with Biserial Correlations ≥ .30	Range of Biserial Correlations	Mean Biserial Correlation	Number of Low X_{50}s (-1.01 or below)	Number of Medium X_{50}s (-1.00 to +1.00)	Number of High X_{50}s (+1.01 or above)	Range of βs	Mean β
7M	246	10	.34-.78	.65	19	.28 to .78	.53	3	17	0	.29 to 1.26	.65
7F	251	9	.50-.75	.63	19	.29 to .71	.53	2	18	0	.31 to 1.00	.65
8M	228	7	.54-.73	.62	19	.24 to .70	.58	2	18	0	.24 to .99	.73
8F	224	9	.56-.77	.65	19	.01 to .84	.59	4	16	0	.01 to 1.54	.80
9M	304	6	.53-.72	.62	20	.30 to .73	.58	1	19	0	.31 to 1.08	.74
9F	302	12	.55-.76	.65	20	.36 to .81	.62	6	14	0	.38 to 1.39	.85
10M	287	13	.59-.71	.66	20	.41 to .83	.67	0	20	0	.46 to 1.46	.96
10F	302	18	.62-.86	.75	20	.48 to .93	.70	9	11	0	.55 to 2.57	1.14
11M	253	17	.61-.81	.72	20	.37 to .90	.70	7	13	0	.40 to 2.04	1.08
11F	265	18	.63-.83	.73	20	.38 to .96	.75	7	13	0	.41 to 3.61	1.35
12M	190	19	.64-.79	.72	20	.40 to .89	.72	10	10	0	.44 to 2.06	1.10
12F	253	17	.60-.86	.75	20	.43 to 1.01	.75	8	11[a]	0	.48 to 8.49[b]	1.59[c]

[a]The X_{50} could not be computed for the item for which the biserial correlation exceeded 1.00.

[b]The highest β could not be computed since the highest biserial correlation exceeded 1.00.

[c]Based on 19 items.

Table 25

Testimony III: Item Statistics for the Total Test
(40 Items)

Grade/Sex	N	Number of Items Within Middle Difficulty Range (.65—.85)	Range of Item Difficulties	Mean Difficulty	Number of Items with Biserial Correlations ≥ .30	Range of Biserial Correlations	Mean Biserial Correlation	Number of Low X_{50}s (-1.01 or below)	Number of Medium X_{50}s (-1.00 to +1.00)	Number of High X_{50}s (+1.01 or above)	Range of βs	Mean β
7M	246	11	.31-.79	.58	32	.15 to .49	.37	8	31	1	.16 to .57	.40
7F	251	15	.31-.75	.59	28	.10 to .60	.36	14	23	3	.10 to .74	.39
8M	227	15	.37-.76	.59	29	.15 to .63	.40	9	29	2	.16 to .82	.44
8F	223	18	.32-.90	.62	28	.05 to .67	.38	17	20	3	.05 to .89	.42
9M	303	18	.33-.77	.59	31	.07 to .64	.41	14	25	1	.07 to .84	.46
9F	305	21	.30-.86	.63	31	-.14 to .69	.41	18	20	2	-.14 to .94	.48
10M	288	23	.37-.82	.64	37	.06 to .74	.47	19	20	1	.06 to 1.10	.55
10F	311	19	.30-.92	.63	34	-.02 to .71	.44	20	17	3	-.02 to 1.00	.52
11M	256	25	.32-.85	.67	37	.04 to .78	.50	21	18	1	.04 to 1.25	.64
11F	262	25	.31-.91	.68	33	-.11 to .76	.49	22	16	2	-.11 to 1.17	.64
12M	195	22	.34-.85	.67	34	.11 to .75	.47	18	21	1	.11 to 1.12	.58
12F	251	20	.29-.92	.68	36	-.05 to .78	.54	21	18	1	-.05 to 1.26	.64

Table 26

Testimony III: Item Statistics for the Recency Subtest
(20 Items)

Grade/Sex	N	Number of Items Within Middle Difficulty Range (.65—.85)	Range of Item Difficulties	Mean Difficulty	Number of Items with Biserial Correlations ≥ .30	Range of Biserial Correlations	Mean Biserial Correlation	Number of Low X_{50}s (-1.01 or below)	Number of Medium X_{50}s (-1.00 to +1.00)	Number of High X_{50}s (+1.01 or above)	Range of βs	Mean β
7M	246	9	.46-.79	.63	18	.15 to .56	.45	5	15	0	.16 to .70	.51
7F	251	13	.42-.76	.63	20	.33 to .62	.47	7	13	0	.35 to .80	.54
8M	228	10	.45-.76	.62	20	.32 to .76	.49	5	15	0	.33 to .91	.58
8F	224	14	.37-.90	.67	19	.28 to .74	.51	8	12	0	.30 to 1.11	.61
9M	304	13	.48-.77	.65	20	.30 to .69	.50	6	14	0	.31 to .95	.61
9F	302	14	.41-.85	.67	19	.08 to .69	.51	10	10	0	.09 to .96	.62
10M	287	15	.48-.82	.71	19	.28 to .77	.55	10	10	0	.29 to 1.24	.67
10F	302	11	.45-.92	.68	19	.23 to .79	.55	13	7	0	.23 to 1.33	.70
11M	253	14	.48-.85	.73	19	.13 to .78	.59	14	6	0	.13 to 1.28	.82
11F	265	14	.50-.91	.74	18	.11 to .83	.59	15	5	0	.11 to 1.21	.85
12M	190	14	.45-.85	.72	19	.26 to .80	.57	12	8	0	.27 to 1.36	.75
12F	253	11	.39-.92	.74	17	.15 to .82	.58	12	8	0	.15 to 1.49	.81

GPO 812-506-3

Table 27

Testimony III: Item Statistics for the Proximity Subtest
(20 Items)

Grade/Sex	N	Number of Items Within Middle Difficulty Range (.65—.85)	Range of Item Difficulties	Mean Difficulty	Number of Items with Biserial Correlations ≥ .30	Range of Biserial Correlations	Mean Biserial Correlation	Number of Low X_{50}s (-1.01 or below)	Number of Medium X_{50}s (-1.00 to +1.00)	Number of High X_{50}s (+1.01 or above)	Range of βs	Mean β
7M	246	3	.31-.74	.55	17	.13 to .58	.42	3	16	1	.14 to .71	.46
7F	251	4	.35-.75	.53	15	.25 to .57	.39	3	15	2	.25 to .69	.45
8M	228	6	.38-.70	.57	19	.22 to .67	.46	3	16	1	.22 to .90	.54
8F	224	6	.33-.82	.49	16	.17 to .71	.42	5	12	3	.17 to 1.02	.48
9M	304	6	.33-.74	.57	19	.24 to .62	.45	5	14	1	.25 to .80	.52
9F	302	7	.30-.80	.59	18	.22 to .67	.47	5	14	1	.22 to .91	.55
10M	287	8	.35-.76	.61	19	.29 to .65	.53	6	13	1	.31 to .86	.63
10F	302	11	.46-.92	.62	19	.23 to .80	.52	6	13	1	.29 to 1.42	.66
11M	253	11	.30-.81	.64	19	.24 to .77	.55	6	14	0	.25 to 1.20	.69
11F	265	9	.32-.83	.65	20	.32 to .82	.56	7	12	1	.33 to 1.45	.73
12M	190	9	.34-.81	.63	19	.17 to .72	.50	5	14	1	.17 to 1.05	.61
12F	253	9	.30-.86	.64	18	.25 to .88	.59	6	13	1	.26 to 1.88	.80

Table 28

Reasoning I: Item Statistics
(28 Items)

Grade/Sex	N	Number of Items Within Middle Difficulty Range (.44—.76)	Range of Item Difficulties	Mean Difficulty	Number of Items with Biserial Correlations ≥ .30	Range of Biserial Correlations	Mean Biserial Correlation	Number of Low X_{50}s (-1.01 or below)	Number of Medium X_{50}s (-1.00 to +1.00)	Number of High X_{50}s (+1.01 or above)	Range of βs	Mean β
7M	245	0	.14-.39	.29	26	.24 to .65	.53	0	14	14	.25 to .87	.63
7F	248	4	.16-.51	.32	26	.23 to .74	.50	1	15	12	.24 to 1.08	.60
8M	230	4	.19-.47	.31	28	.32 to .73	.56	0	19	9	.33 to 1.07	.69
8F	218	11	.18-.62	.36	28	.32 to .81	.59	0	23	5	.34 to 1.37	.74
9M	302	4	.21-.50	.34	28	.33 to .69	.57	0	21	7	.35 to .96	.71
9F	301	10	.22-.65	.40	28	.33 to .86	.62	0	24	4	.36 to 1.66	.82
10M	277	16	.31-.66	.41	28	.43 to .80	.62	0	28	0	.48 to 1.32	.85
10F	294	20	.35-.89	.51	28	.50 to .90	.69	0	28	0	.58 to 2.04	1.02
11M	262	23	.33-.69	.49	28	.46 to .84	.69	0	28	0	.51 to 1.55	1.02
11F	270	23	.32-.72	.53	28	.47 to .88	.69	1	27	0	.54 to 1.85	1.02
12M	191	22	.30-.66	.47	28	.42 to .86	.71	0	28	0	.47 to 1.67	1.07
12F	252	23	.36-.79	.55	28	.41 to .86	.68	1	27	0	.45 to 1.69	.97

Table 29

Reasoning II: Item Statistics
(28 Items)

Grade/Sex	N	Number of Items Within Middle Difficulty Range (.65—.85)	Range of Item Difficulties	Mean Difficulty	Number of Items with Biserial Correlations ≥ .30	Range of Biserial Correlations	Mean Biserial Correlation	Number of Low X_{50}s (-1.01 or below)	Number of Medium X_{50}s (-1.00 to +1.00)	Number of High X_{50}s (+1.01 or above)	Range of βs	Mean β
7M	246	4	.40-.77	.60	20	.10 to .60	.38	8	18	1	.10 to .76	.42
7F	251	10	.28-.79	.62	23	-.08 to .68	.41	8	19	1	-.08 to .93	.47
8M	227	12	.40-.76	.64	25	.10 to .55	.43	9	17	1	.11 to .66	.49
8F	223	14	.32-.82	.64	25	-.05 to .69	.45	12	15	1	-.05 to .96	.54
9M	303	16	.42-.80	.63	25	.26 to .65	.48	7	21	0	.27 to .87	.57
9F	305	18	.36-.86	.69	27	.07 to .72	.51	13	14	1	.07 to 1.03	.62
10M	288	23	.41-.83	.69	26	.28 to .74	.55	16	11	0	.29 to 1.11	.68
10F	311	19	.36-.90	.59	24	.08 to .81	.60	19	8	1	.08 to 1.37	.82
11M	256	24	.51-.87	.79	27	.24 to .82	.60	16	12	0	.24 to 1.42	.79
11F	262	17	.37-.93	.74	27	.28 to .85	.60	22	6	0	.29 to 1.61	.81
12M	195	24	.49-.87	.74	28	.35 to .77	.60	18	10	0	.37 to 1.21	.79
12F	251	16	.31-.94	.77	28	.31 to .86	.64	21	6	1	.34 to 1.72	.90

Table 30

Reasoning III: Item Statistics
(28 Items)

Grade/Sex	N	Number of Items Within Middle Difficulty Range (.47—.78)	Range of Item Difficulties	Mean Difficulty	Number of Items with Biserial Correlations ≥ .30	Range of Biserial Correlations	Mean Biserial Correlation	Number of Low X_{50}s (-1.01 or below)	Number of Medium X_{50}s (-1.00 to +1.00)	Number of High X_{50}s (+1.01 or above)	Range of βs	Mean β
7M	245	11	.23-.64	.43	24	.15 to .74	.51	0	22	6	.16 to 1.11	.64
7F	248	14	.25-.75	.48	25	-.09 to .71	.50	2	23	3	-.09 to 1.01	.57
8M	230	16	.16-.70	.49	27	.07 to .80	.56	0	27	1	.07 to 1.36	.73
8F	218	22	.18-.81	.55	26	.18 to .82	.57	1	25	2	.18 to 1.45	.73
9M	302	19	.16-.72	.49	27	.22 to .78	.59	0	26	2	.23 to 1.25	.78
9F	301	20	.15-.76	.57	27	.03 to .77	.58	5	22	1	.03 to 1.19	.73
10M	277	22	.23-.81	.63	27	.29 to .87	.63	8	19	1	.30 to 1.75	.87
10F	294	19	.20-.89	.67	27	.02 to .83	.60	12	15	1	.02 to 1.48	.82
11M	262	22	.26-.83	.66	27	.28 to .83	.66	5	22	1	.29 to 1.48	.93
11F	270	16	.25-.90	.72	28	.33 to .92	.64	12	15	1	.35 to 2.38	.90
12M	191	24	.30-.83	.64	28	.38 to .88	.68	1	26	1	.41 to 1.93	1.02
12F	252	16	.25-.92	.72	28	.32 to .85	.61	16	11	1	.33 to 1.63	.84

Table 31

Reasoning IV: Item Statistics
(28 Items)

Grade/Sex	N	Number of Items Within Middle Difficulty Range (.53—.80)	Range of Item Difficulties	Mean Difficulty	Number of Items with Biserial Correlations ≥ .30	Range of Biserial Correlations	Mean Biserial Correlation	Number of Low X_{50}s (-1.01 or below)	Number of Medium X_{50}s (-1.00 to +1.00)	Number of High X_{50}s (+1.01 or above)	Range of βs	Mean β
7M	245	15	.32-.76	.52	25	-.07 to .64	.47	2	24	2	-.07 to .84	.54
7F	248	17	.27-.78	.54	25	.09 to .71	.45	5	20	3	.09 to 1.01	.52
8M	230	17	.21-.81	.57	26	.09 to .76	.46	5	21	2	.09 to 1.18	.55
8F	218	18	.26-.83	.59	25	-.02 to .70	.48	7	20	1	-.02 to .99	.57
9M	302	18	.20-.83	.58	26	.17 to .72	.51	5	22	1	.17 to 1.05	.63
9F	301	21	.28-.87	.63	26	.10 to .78	.53	18	9	1	.10 to 1.23	.67
10M	277	21	.27-.87	.66	25	.23 to .83	.55	10	17	1	.24 to 1.51	.70
10F	294	19	.32-.91	.70	25	.24 to .77	.52	17	10	1	.24 to 1.20	.66
11M	262	18	.48-.88	.68	26	.25 to .79	.55	14	13	1	.26 to 1.29	.71
11F	270	18	.44-.90	.60	28	.34 to .77	.52	20	7	1	.36 to 1.20	.65
12M	191	20	.30-.83	.67	28	.30 to .82	.60	9	18	1	.32 to 1.41	.81
12F	252	15	.35-.94	.78	25	.17 to .87	.51	21	6	1	.17 to 1.78	.66

VI
CONCLUSIONS

The reliability estimates obtained for all of the tests for each age and sex group are sufficient for research purposes and to evaluate group differences. In addition, for some of the tests, particularly for Grades 10–12, the reliability estimates are of a sufficient magnitude to allow for evaluation of differences among individuals. If the further study of the dimensionality of the tests indicates that the subtests of Testimony I and Testimony III should be considered as independent tests they should be lengthened to be more reliable.

The items, in general, exhibit the characteristics sought by the investigators. Many of the items fall within the middle difficulty range. Most items discriminate rather sharply, as indexed by high biserial correlations and ßs. Most of the items which have low biserial correlations and ßs are found in one of two tests, Testimony I or Testimony III, when total test score is the criterion measure. These low correlations may be indications that at least some items are measuring different abilities and that subtests should perhaps be retained. Most of these same items have correlations and ßs above .30 for the appropriate subtest when it is the criterion measure. As evidenced by the X_{50} item statistics, many more items are maximally discriminating among students of low and middle abilities than among students of high ability. Thus, these items are discriminating more clearly among less able students than they are among more able students. In general, the item statistics tend to increase in value from Grade 7 to Grade 12.

Although the final edition of the tests was designed primarily for Grades 10–12, there are indications that the tests might also yield useful information for Grades 7–9. A more exact interpretation of the adequacy of the reliability and item statistics of the tests is left to the reader and potential user who should judge the value of the tests for his particular purpose.

REFERENCES

Allen, R. R., J. D. Feezel, and F. J. Kauffeld. A taxonomy of concepts and critical abilities related to the evaluation of verbal arguments. Occasional Paper Number 9. Wisconsin Research and Development Center, The University of Wisconsin, Madison, August 1967.

Allen, R. R., J. D. Feezel, and F. J. Kauffeld. The Wisconsin tests of testimony and reasoning assessment (WISTTRA). Practical Paper Number 6. Wisconsin Research and Development Center for Cognitive Learning, The University of Wisconsin, Madison, December 1968.

Baker, Frank B. Origins of the item parameters X_{50} and β as a modern item analysis technique. Journal of Educational Measurement, 1965, 2, 167-180.

Baker, Frank B. Test analysis package: A program for the CDC 1604-3600 computers. Department of Educational Psychology, The University of Wisconsin, June 1966.

Baker, F. B. and T. J. Martin. Fortap: A fortran test analysis package. Wisconsin Research and Development Center for Cognitive Learning, The University of Wisconsin, 1968.

Dale, Edgar, and Jeanne S. Chall. A formula for predicting readability: Instructions. Educational Research Bulletin, 1948, 27, 37-54.

Ebel, Robert L. Measuring educational achievement. Englewood Cliffs, N. J.: Prentice-Hall, Inc., 1965.

Harris, Margaret L. A factor analytic study of the Wisconsin tests of testimony and reasoning assessment (WISTTRA). Wisconsin Research and Development Center for Cognitive Learning, The University of Wisconsin, Madison, 1969 (in progress).

Miller, Donald M., et al. Multivariate procedures for stratifying school districts. Laboratory of Instructional Research, The University of Wisconsin, Madison, 1967.

Rott, Robert K., Jerry D. Feezel, and R. R. Allen. A study of student abilities in the evaluation of verbal arguments. Wisconsin Research and Development Center for Cognitive Learning, The University of Wisconsin, Madison, in press.

Thorndike, Edward L. and Irving Lorge. The teacher's wordbook of 30,000 words. Columbia: Teachers College Press, 1944.

Thorndike, Robert L. Reliability. In E. F. Lindquist (Ed.), Educational measurement. Washington, D. C.: American Council on Education, 1951.

Thorndike, Robert L. and Elizabeth Hagen. Measurement and evaluation in psychology and education. New York: John Wiley & Sons, Inc., 1961.

Toulmin, Stephen. The uses of argument. Cambridge: University Press, 1958.

GPO 812-508-2